Journal of Mathematical Psychology 57 (2013) 177–179

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Commentary

On biases in assessing replicability, statistical consistency and publication bias

Valen E. Johnson

Department of Statistics, 464C Blocker Building, 3143 TAMU, College Station, TX 77843-3143, United States

HIGHLIGHTS

- Biases associated with the selection of test statistics to detect publication bias are described.
- Problems with the failure to identify the sampling space of articles are discussed.
- Issues regarding the interpretation of significance tests in journals are presented.

ARTICLE INFO

Article history:
Received 22 April 2013
Available online 14 June 2013

Keywords:
Significance tests
Bayes factors
Publication bias
Excess of significant findings
Uniformly most powerful Bayesian tests

ABSTRACT

Methodology described by Francis in "Replication, Statistical Consistency and Publication Bias" is examined in the context of its application to the Topolinski and Sparenberg (2012) article. Several biases are discovered in this analysis, including selection biases in the reporting of *p*-values from the Topolinski and Sparenberg article, as well as in the criteria that were used in the selection of this article. General concerns regarding the replicability of scientific studies based on significance tests conducted at the 5% level of significance are also described.

My wife and I recently attended a viewing of the blockbuster movie *Skyfall*. My wife, who is not much a James Bond fan, cleverly pointed out at the end of the initial chase scene that there was really no way Bond could have survived the fall from the bridge. I was a bit dumbfounded by her comment, coming as it did after Bond had already driven his motorcycle across the rooftops of Istanbul, catapulted himself from a motorcycle on to the top of a moving train, and used a backhoe to both deflect a barrage of machine gun fire and recouple two segments of a train. Clearly, she does not understand the concept of "willing suspension of disbelief" (Coleridge, 1817).

I find myself in much the same position after being asked to comment on Francis's article in which a testing framework is proposed for testing for an excess of significant findings in the psychological literature. I simply cannot quite determine the level of absurdity that I am expected to ignore. It is almost as if all parties involved are pretending that *p*-values reported in the psychological literature have some well-defined meaning and that our goal is to ferret out the few anomalies that have somehow misrepresented a type I error. Nothing, of course, could be farther from the truth.

Before discussing the details of the article, I think it is perhaps worthwhile to review a fundamental truth of classical statistical hypothesis testing and the report of *p*-values. That truth is this: as normally reported, *p*-values and significance tests provide the consumer of these statistics absolutely no protection against rejecting "true" null hypotheses at less than any specified rate smaller than 1.0. *P*-values and classical significance tests only provide the experimenter with such a protection. And they only provide an experimenter with this protection if she behaves in a scientifically principled way. If you do not agree, consider the following stylized example in the realm of medical research.

Suppose that a biological pathway associated with the growth of cancer is identified and that numerous teams of medical researchers develop drugs to disrupt this pathway. Assume that 1,000 such drugs are developed and tested in 5% significance tests. Suppose further that all tests are conducted in a scientifically valid way – that is, only the primary outcome of each experiment is analyzed, no sub-group analyses are performed, no anomalous findings are discarded, etc. – and that only those drugs found to be significant at the 5% level are reported and published.

What would the result of this activity be if it was later determined that the identified biological pathway had nothing at all to do with the development of cancer? On average, 50 drugs would have been identified as promising. If these drugs were subjected to further independent testing at the 5% level of significance, on average 2.5 of these drugs would have had their effects confirmed in follow-up experiments. According to statistical theory, we know that the researchers will commit, on average, a type I error in only

E-mail address: vjohnson@stat.tamu.edu.

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.04.003

5% of the trials. But what is the type I error rate of the journals that reported the significant findings? It is 100%—the null hypothesis of no treatment effect is, by construction, true for all tested drugs. The same is true for the reports of significance for the average 2.5 drugs that would have passed the replication studies.

Note also that this stylized example does not account for the fact that in most scientific experiments there are multiple outcome variables or that researchers often perform multiple statistical analyses of each outcome variable in order to report the most highly significant result. Or that researchers frequently terminate experiments prematurely when a significant test statistic is obtained or that observations not compatible with an investigator's hypothesis are sometimes discarded because of "errors" in data collection.

One might argue that the solution to this problem is to simply register all experiments and their statistical analysis plans prior to their execution. This is essentially the intent of the Food and Drug Administration's effort to register clinical trials. Unfortunately, the reality of scientific research in most areas of study is that there are no comprehensive databases for preregistration of experiments. As a consequence, it should come as no surprise that published research findings, particularly of novel findings, often fail to replicate in follow-up studies.

Part of the problem of non-reproducibility of scientific studies can be attributed to the declaration of statistical significance for experiments that have less than a 5% type I error. For z tests and other uniformly most powerful tests, the rejection regions associated with this level of significance correspond exactly to the rejection regions of uniformly most powerful Bayesian tests (Johnson, 2013). From this correspondence, it can be shown that a p-value of 0.05 reduces the odds that the null hypothesis is true by a factor of less than 4. To obtain evidence that increases the odds of the alternative hypothesis by a factor of 20, significance tests must be conducted at the 0.7% level. Requiring p-values to be less than 0.5% for a claim of significance would certainly be a step in the right direction.

Of course, even this improvement would not eliminate the problem. In the drug example above, on average 5 drugs would have passed initial testing at this more stringent criterion. It is for this reason that science must proceed by replication of studies. Once a finding has been published, it then becomes feasible for journals to publish additional studies that both confirm and deny the original finding. This process can, of course, be facilitated through the report of Bayes factors, but that is the topic for another paper.

Francis proposes to address the problem of non-replicability of scientific studies through the implementation of a test for an excess in significant findings. Although I applaud him for this effort, I have concerns regarding the statistical methodology that he employs, as well as concerns regarding his guilt in committing sins of the very type he hopes to expose.

The type of publication bias that Francis hopes to address concerns the publication of results in which p-values from several experiments are all much closer to the nominal level of, say, 5% than would be expected under usual random variation. Francis points out that the appearance of such closely spaced p-values could result either from an investigator's failure to report test results that were not significant, or, more insidiously, the fabrication of data. Following previous research along these lines (e.g., Ioannidis & Trikalinos, 2007), he proposes a chi-squared test statistic for an excess of significant findings that is based on calculating the observed and expected number of significant findings under the assumption of a common (or nearly common) effect size across studies. As he shows through extensive simulation studies, the actual type I error of the resulting test statistic falls far below its nominal level. Francis correctly argues that the resulting test is conservative, but the disparity between the advertised type I error rate and its actual rate raises serious questions about the veracity of the entire procedure. Clearly, a test statistic that more nearly achieved its nominal operating characteristic would be easier to interpret. It would also provide better power in detecting the type of publication bias that Francis targets.

Aside from the fact that the operating characteristics of Francis's test statistic are so poor, even more serious problems are encountered when one attempts to apply this methodology. These problems are perhaps easiest to understand by reviewing Francis's analysis of the Topolinski and Sparenberg (2012) article.

Topolinski and Sparenberg (T&S) conducted four studies to determine whether subjects modified their responses to stimuli according to the direction in which another object was rotated. In their first study, T&S randomized 50 subjects into two groups. One group turned objects in the clockwise direction while rating the likeness of Chinese ideographs, whereas the other turned objects in the counterclockwise direction while also rating the likeness of ideographs. Likeness was rated on a 7 point scale. During a training phase, both groups were shown 10 Chinese ideographs. In the test phase, both groups were shown the same 10 ideographs, plus 10 "new" ideographs, again while turning objects. The parameter of interest in this experiment was the difference, between the two groups, of the mean difference assigned by each subject to ratings of the original and new ideographs. T&S hypothesized that the mean difference of differences would be affected by the direction in which subjects turned the objects.

T&S reported several statistics based on data collected in Study 1 to support their hypothesis. They begin their results section as follows:

> A 2 (Exposure: old items, new items, within) × 2 (Turning Direction: clockwise, counterclockwise; between) analysis of variance (ANOVA) solely yielded an interaction between Exposure and Turning Direction, $F(1, 48) = 15.93$, $p < .0001$, $\eta_p^2 = .25$, and no other effects (ps $< .30$). Participants who had turned the objects counterclockwise liked old stimuli ($M_{counterclockwise\text{-}old} = 4.46$, $SD = 0.59$) more than novel stimuli ($M_{counterclockwise\text{-}new} = 4.10$, $SD = 0.57$), $t(24) = 3.64$, $p < .001$, $d = 0.62$, replicating the classic mere exposure effect. In contrast, however, participants who turned the objects clockwise liked old stimuli ($M_{counterclockwise\text{-}old} = 3.95$, $SD = 0.86$) less than novel stimuli ($M_{counterclockwise\text{-}new} = 4.27$, $SD = 0.85$), reversing the mere exposure effect, $t(24) = 2.30$, $p < .031$, $d = 0.36$.

From this description, it is clear that T&S regard their primary outcome variable as the difference between the difference ratings provided by subjects who rotated the objects clockwise versus those who turned the objects counterclockwise. The test statistic for this difference of differences corresponds to the initial F statistic for the interaction term in the ANOVA, $F(1, 48) = 15.93$. The corresponding t statistic is $t(48) = 3.99$, is much larger than either of the t statistics reported in the secondary analyses (3.64 and 2.30). That is, the interaction, and not the simple main effects, is the primary test statistic. It must be emphasized that both t-statistics that Francis includes in his analysis represent tests of whether a difference between the original and novel ideograph ratings under the same experimental condition (either clockwise or counterclockwise rotation) was equal to 0. However, the appropriate analysis for this experiment is to test whether the difference between these differences is 0. Because the differences have opposite signs, the appropriate analysis results in a much larger estimate of the experiment's effect size and statistical power. Note also that there is only one (not two) significant finding in this study.

Francis's error in this example signals an important practical problem associated with implementing tests for an excess of

significant findings. Most journal articles report several analyses for every data set reported, and in many cases it is not clear which analysis and test statistics should be selected to test for an excess of significant findings. In this case the F statistic is clearly correct, but in many studies the choice of test statistic is not nearly so clear.

In Study 2 of T&S, subjects were again randomized to turn an object in one of two directions, but in this experiment subjects completed an experiential openness survey. T&S report that the mean response to survey items representing "openness" for subjects who turned the object counterclockwise was higher than those who turned the object counterclockwise, with $t(58) = 2.21$ and $p = .031$. A somewhat more significant p-value is obtained through an analysis that controlled for mood and arousal ($F(1, 56) = 6.54$, $p = 0.013$), but this result does not appear to represent the analysis that was planned prior to the experiment. Interestingly, T&S report that no significant finding is discovered for mood and arousal ratings, but that no significant finding was anticipated for these attributes. This explanation might represent "HARKing," but it does suggest that T&S have not intentionally placed unfavorable outcomes in a "file drawer".

Errors similar to Study 1 are made in the interpretation of results from Study 3, which essentially replicates Study 1 and Study 2, except that subjects in Study 3 watched a rotating object rather than actually rotating an object themselves, and the dependent variables of both Study 1 and 2 were measured in counterbalanced order. The F statistic in Study 3 that is similar to the F statistic used in Study 1 is $F(1, 79) = 9.54$, with $p = .003$. The t statistic corresponding to Study 2 is $t(79) = 2.04$ and $p < .044$. Francis decides to base his analysis of an excess of statistical findings on the weaker t statistic and ignores the stronger result based on the F statistic, since, as he puts it, "these additional tests can only decrease the power of the overall findings (the rules of probability dictate that multiple outcomes in a set cannot be more probable than a single outcome from that set)".

Study 4 reports the outcome from only a single test, which, as Francis reports, produced a t statistic of $t(48) = 2.04$ and $p = .047$.

What is the impact of Francis' selection of test statistics on tests for excess significant findings? The power associated with the F statistic in Study 1 is approximately 0.97 when calculated using methods similar to those proposed for the other statistics included in the study. Francis reports the power of Studies 2 and 4 to be 0.573 and .503, respectively. The power of the second finding in Study 3 is reported as .514, and the power of the first part of Study 3 is approximately 0.86. According to the proposed test statistic and Eq. (3), the probability of seeing 5 significant findings out of 5 is thus $0.97 \times .573 \times .86 \times .513 \times .503 = .12$, which is not significant according to the criterion described in the article. It should be noted that this calculation does not account for the uncertainty in the power values, or the fact that the use of a pooled (averaged) power value would have led to an even less significant result. The errors made by Francis in interpreting the results of the T&S article are, unfortunately, propagated into his post-hoc test analysis, invalidating that analysis as well (the post-hoc testing paradigm itself seems even more ad hoc; it relies on numerous unverified and untestable assumptions that, for reasons of time and space constraints, are not discussed further here).

Several lessons are apparent from the preceding discussion. As mentioned above, it is often difficult to determine which test statistics should be used to test for an excess of statistical findings. Most quantitative research articles report multiple tests based on the same data. The resulting test statistics are generally not independent, which complicates joint modeling of their values. And, as demonstrated above, the choice of test statistics to use in a test for excess significant findings can have a dramatic effect on the conclusions of the test. In addition, statistical power typically varies substantially across distinct experiments. Although Francis combined results within the T&S studies so that the power associated with each study ranged between 0.503 and 0.573, the actual powers of the individual experiments reported by T&S appear to have ranged from about 0.5 to 0.97. Heterogeneous power values further complicate the calculation and interpretation of the test statistic in Eq. (1), as Francis himself notes. Similarly, the assumption that standardized effect sizes are homogeneous across experiments should be regarded with caution. In the T&S article, for example, there does not seem to be any rationale to justify an assumption that the standardized effect size of rotation on object likeness in Study 1 should be the same as the standardized effect size of rotation on the mean response to survey items, as reported in Study 2. In general, different experiments measure different outcomes under different conditions, so it will seldom be the case that a standardized effect size of a common treatment is the same across experiments. It therefore seems that tests of the type proposed by Francis are unlikely to find broad application in practical settings.

As serious as these problems are, I think a much more serious deficiency of Francis' program involves exactly the type of publication bias that Francis is attempting to detect. As the drug example above demonstrates, it is impossible to evaluate the type I error associated with Francis' report of excess significant findings because there is no way to evaluate his sampling frame. How many articles did Francis examine before finding four similarly-valued t statistics in the T&S paper? And how many t statistics did he examine in each of those papers? Despite his protestations to the contrary, his methodology is subject to exactly the same type of publication bias that he attempts to expose.

Additional references

Coleridge, S. T. (1985). *Biographia literaria: the collected works of Samuel Taylor Coleridge*. Princeton University Press.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253.

Johnson, V.E. (2013). Uniformly most powerful Bayesian tests, To appear in the Annals of Statistics.

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychological and Personality Science*, *3*, 308–314.